Central European Science Journals
www.cesj.com

Central European Journal of Chemistry
Central European Science Journals
CEJC 3(4) 2005 731–741

SVD-based principal component analysis of geochemical data

Petr Praus*

Department of Analytical Chemistry and Material Testing,
VSB-Technical University Ostrava,
17. listopadu 15, 708 33 Ostrava, Czech Republic

Received 2 May 2005; accepted 20 June 2005

Abstract: Principal Component Analysis (PCA) was used for the mapping of geochemical data. A testing data matrix was prepared from the chemical and physical analyses of the coals altered by thermal and oxidation effects. PCA based on Singular Value Decomposition (SVD) of the standardized (centered and scaled by the standard deviation) data matrix revealed three principal components explaining 85.2 % of the variance. Combining the scatter and components weights plots with knowledge of the composition of tested samples, the coal samples were divided into seven groups depending on the degree of their oxidation and thermal alteration.
The PCA findings were verified by other multivariate methods. The relationships among geochemical variables were successfully confirmed by Factor Analysis (FA). The data structure was also described by the Average Group dendrogram using Euclidean distance. The found sample clusters were not defined so clearly as in the case of PCA. It can be explained by the PCA filtration of the data noise.

Keywords: Principal component analysis, singular value decomposition, factor analysis, hierarchical cluster analysis, altered coals, geochemistry

1 Introduction

Real chemical data contain not only important information but also confusing noise. They are mostly far from normality with collinear and/or autocorrelated variables, containing outliers and so forth. For extraction of this information with its minimal lost, there are several chemometric methods for the reduction of data dimensionality, such as Principal Component Analysis, Factor Analysis, Independent Component Analysis [1], Independent

* E-mail: petr.praus@vsb.cz

Factor Analysis [2], Generative Topographic Mapping [3], etc.

In PCA, we look for new abstract orthogonal components (eigenvectors) which explain the most of data variation. PCA is based on the Eigenvalue Decomposition (EVD) [4,5] of covariance/correlation matrixes and/or by the SVD of real data matrices. In comparison with EVD, SVD is the more robust, reliable, and precise method with no need to compute the input covariance/correlation matrix [6]. From a numerical point of view, SVD is well known for its stability and convergence, even in cases of ill-conditioned problems.

SVD decomposes an arbitrary matrix A (nxp) into three matrices:

$$\mathrm{A} = \mathrm{U}\ \mathrm{S}\ \mathrm{V}^T \tag{1}$$

where U (nxn) and V^T (pxp) are orthogonal and normalized matrices, i.e., $\mathrm{U}^T\mathrm{U}=\mathrm{I}$ and $\mathrm{V}^T\mathrm{V}=\mathrm{I}$. S (nxp) is a diagonal matrix with singular values in decreasing order. The columns of U are the left singular vectors and the rows of the V^T are the right singulars vectors. Computing the SVD consists of finding the eigenvalues and eigenvectors of AA^T and $\mathrm{A}^T\mathrm{A}$. The columns of U are eigenvectors of AA^T and the rows V^T are the eigenvectors of $\mathrm{A}^T\mathrm{A}$. The singular values are the square roots of the eigenvalues of AA^T or $\mathrm{A}^T\mathrm{A}$. The powerful property of SVD is compressing the information contained in A into the first singular vectors which are mutually orthogonal and their importance rapidly decreases after the first columns/rows. The importance of each singular vector is given by the squares of nonnegative singular values of the matrix S.

SVD has already found the wide range of various applications in molecular dynamic and gene expression analysis [7], information retrieval, e.g., in the technique of Latent Semantic Indexing [8], image processing [9], spectral analysis [10], etc.

The aim of this paper was to analyze geochemical data by the SVD-based PCA (PCA/SVD). The testing data matrix summarises the results of chemical and physical properties of the coal samples taken from the Upper Silesian Coal Basin in the Czech Republic. There are Carboniferous red bed bodies in this area. Within the red beds and their vicinity, coals were altered by the oxidation and thermal effects which are manifested in their macroscopic and microscopic characteristics and various chemical composition.

2 Experimental

2.1 Geochemical data set

PCA/SVD was tested on the data matrix of the coal samples (n= 52) taken from the region of red bed bodies of the Upper Silesian Coal Basin. This testing data set was adopted from the work of Klika and Kraussová [11] with the authors courtesy. Coal analyses, including sampling and preservation, were carried out according to the standard methods. The parameters were selected in order to classify altered coals. The content of ash (A^d, wt %), moisture (W^a, wt %), volatile matter (V^{daf}, wt %), humic acids (HA^{daf}, wt %), combustion heat (Q^{daf}, MJ/kg), mean reflectance of vitrinite (R_O), concentration of elements (C^{at}, H^{at}, O^{at}, N^{at}, all in atom %). The sample summary statistics are given

in Table 1. It is obvious that the data are not normally distributed and scaling effects can be expected.

2.2 Multivariate computations

The data matrix for PCA/SVD was prepared and treated in Excel 97. SVD of this matrix A was executed using the standard MATLAB command svds(A,k) which computes the k largest values and associated singular vectors of the matrix A. These data are typical by various types and scales of measured variables. To avoid the scaling effects, the data were standardized, i.e. centered the mean (average) and scaled the standard deviation of the original measurement variables [12]. FA and HCA were performed by the software packages NCSS 97 (Number Cruncher Statistical Systems, Utah).

3 Results and discussion

3.1 Determination of the principal components

Determination of the components number is given by the characteristics of singular values and is demonstrated in a scree plot (see Fig. 1). The singular values sharply decrease within three largest singular vectors and then slowly stabilize for remaining ones which contain a great deal of noise and therefore are not useful. Regarding the SVD theory, the singular values correspond to the square roots of the eigenvalues. That is why the variance of the singular vectors (principal components, PCs) can be expressed according to the equation

$$var. = \frac{s_k^2}{\sum_1^n s_i^2} \tag{2}$$

where s_k is a singular value. The revealed PC1 to PC3 contain 46.2 %, 27.4 %, and 11.6 %, i.e. 85.2 % of the total data variance. This is in close agreement with the conventional 80% of the variance which should be explained by the principal components.

3.2 PCA of the relationships among geochemical variables

The components weights were calculated as the correlation coefficients of the original variables with the principal components and their plot is displayed in Fig. 2. The strong relations among R_o, C^{at}, and their reciprocity to H^{at} are evident from their opposite positions in the plot and well agree with the geochemical theory. The reduction of H^{at} is associated with the relative increase of C^{at}. The high reflectance of vitrinite R_0 is an effect of the coal oxidation at higher temperatures. The chemical composition of thermally altered coals without an influence of oxygen is characterized by the high concentrations of hydrogen. Under these conditions, R_o, C^{at}, and H^{at} can express the intensity of the thermal changes in coals.

Parametr	W^a (%)	A^d (%)	V^{daf} (%)	Q^{daf} (MJ/kg)	HA^{daf} (%)	R	C^{at} (%)	H^{at} (%)	N^{at} (%)	O^{at} (%)
Count	52	52	52	52	52	52	52	52	52	52
Average	10.6	14.7	24.1	30.4	13.3	2.07	64.7	26.1	1.41	7.598
Variance	47.0	152	135	14.06	712	1.86	71.5	60.2	0.728	17.1
Stnd. dev.	6.86	12.3	11.6	3.75	26.7	1.362	8.46	7.76	0.8529	4.14
Minimum	1.07	2.83	2.28	23.44	0*	0.88	55.47	9.34	0.71	1.72
Maximum	24.65	55.51	45.70	35.50	88.41	5.70	88.06	40.32	4.43	15.27
Range	23.58	52.68	43.42	12.06	88.41	4.82	32.59	30.98	3.72	13.55
Skewness	0.7214	5.3599	-1.3122	-1.0423	5.4920	4.1900	3.2678	0.4316	6.2286	0.7645
Kurtosis	-1.6016	4.2330	-1.2066	-1.7067	2.8191	1.2955	0.9257	-0.3927	6.1052	-1.8078

* Concentration bellow the detection limit;
Stnd. dev. = Standard deviation.

Table 1 Summary statistics of the tested geochemical data.

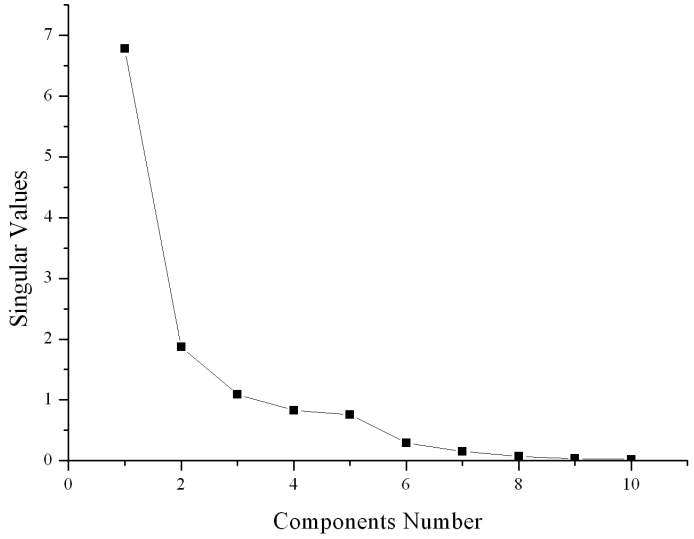

Fig. 1 Scree plot of the singular values.

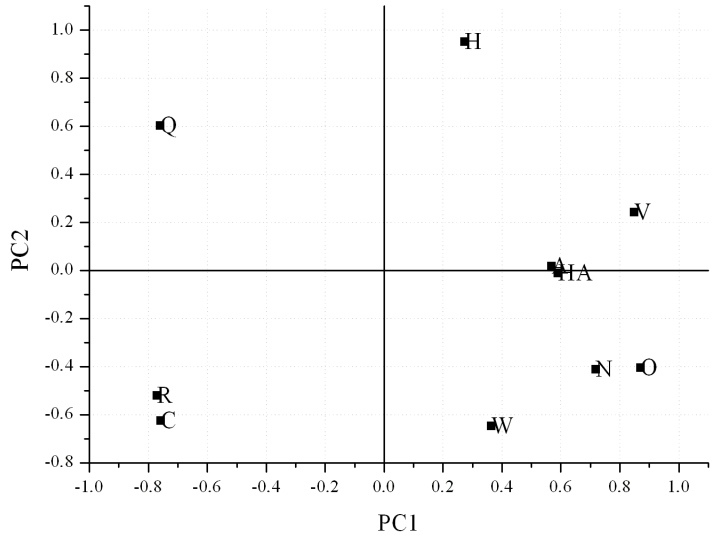

Fig. 2 Principal components weights plot of the geochemical variables.

On the other hand, the variables W^a, O^{at}, HA^{daf}, and A^d are considered to indicate the oxidation processes in coal. Their positions are similar to each other in Fig. 2. In addition, the closest relation between HA^{daf} and A^d is exhibited. The high content of A^d is likely caused by loses of the gaseous oxidation products, such as CO, CO_2, and H_2O. Humic acids and water are the direct products of oxidation processes in altered coals. The reciprocal relationships of W^a and O^{at} to Q^{daf} are obvious from Fig. 2 and can be logically expected.

3.2.1 Factor Analysis of the geochemical data

Factor Analysis was carried out to confirm the relations among the variables found by PCA/SVD. The loading factors are given in Table 2. In factor 1, the content of H^{at} strongly, negatively correlates with R_o and C^{at}. As it was given above, these variables can characterize the thermal coal alteration with zero or very low content of oxygen. Factor 2 reveals the significant correlations among the content of W^a and O^{at} which are negatively correlated with Q^{daf} and thus corresponds to the oxidation alteration of coals. Factor 3 is mainly saturated by the content of A^d and HA^{daf} as the products of coal oxidation, as well. Their relation was clearly shown in Fig. 2.

Parameter	Factor 1	Factor 2	Factor 3
W^a	-0.0538	**0.9080**	-0.2063
A^d	0.1657	0.0301	**0.8205**
V^{daf}	0.7771	0.4159	0.2427
Q^{daf}	-0.0510	**-0.8573**	-0.4502
HA^{daf}	0.1948	0.1139	**0.7430**
R_o	**-0.9364**	-0.19130	-0.1503
C^{at}	**-0.9487**	-0.0361	-0.2720
H^{at}	**0.8728**	-0.4698	-0.0081
N^{at}	0.0734	0.5812	0.6348
O^{at}	0.2857	**0.8197**	0.4268

Table 2 The factor loadings after Varimax rotation.

On the basis of these results, it can be concluded that the relationship among variables within all three factors well agree with those discovered by PCA using the components weights plot.

3.3 PCA clustering of the geochemical data

The principal component scatter plot of PC2 vs. PC1 was constructed (Figs. 3). The seven clearly separated groups were found. The coal samples were divided into these groups by means of hierarchical clustering of the two largest principal components scores. A suitable clustering method was chosen according to the three clustering criteria, such as Cophenetic correlation coefficient (CC), Delta(0.5), and Delta(1.0) [13]. Applying the Euclidean distance metric, the highest CC and the lowest Delta(0.5) and Delta(1.0) were obtained for the Simple Average method (CC=0.7113, Delta(0.5)=0.2855, and Delta(1.0)= 0.3369). The sample groups in Fig. 3 were created exactly according to the clusters of the PC1 and PC2 dendrogram. By looking at this plot, the sample CM-2 seems to be an outlier within the group I (see below). It is caused by its atypical content of V^{daf} (19.98 %) and R_o (2.80) which were proved by the Dean-Dixon test to be the outlying values.

On the basis of the diagnostic plots and the composition of the sample groups I to IV/2 (Table 3), it is evident that the groups are vertically and horizontally located with respect

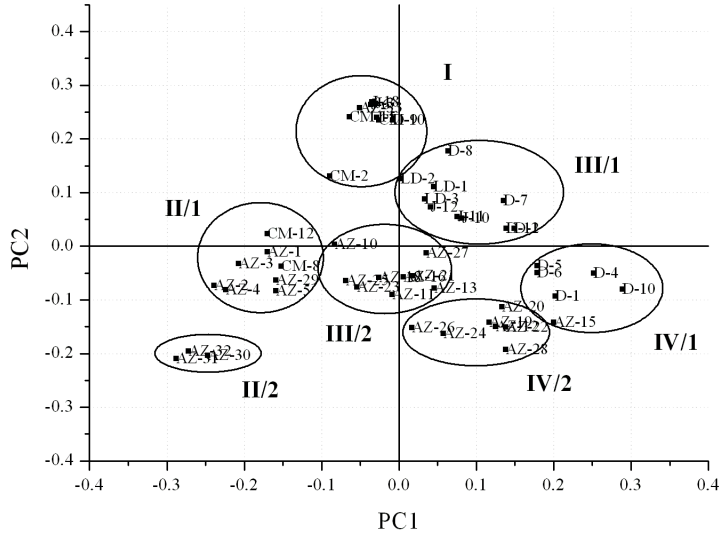

Fig. 3 Principal components scatter plot of the coal samples.

to the thermal and oxidation changes during coal alteration, respectively. The group of samples with the highest concentrations of hydrogen, the lowest values of the vitrinite reflectances and the lowest concentrations of water was denoted as I. These samples are considered to be mainly thermally altered. Six other groups contain the samples altered both thermally and by oxidation and therefore they can be divided into two subgroups (1 and 2) according to the temperatures (lower and higher) of coal oxidation: II (1, 2), III (1, 2), and IV (1, 2). The samples II (1, 2) have the highest content of C^{at} and R_0. The magnitudes of the variables W^a, O^{at}, HA^{daf} indicating oxidation processes increase in the order II/1, III/1, and IV/1. Moreover, the content of W^a is always higher in the subgroups 2 (at higher temperatures). The groups III/1 and IV/1 are significant by the high content of HA^{daf}. The lower content of HA^{daf} in III/2 and IV/2 is likely caused by their thermal decomposition at the temperatures above 250 °C [14].

3.3.1 Hierarchical Clustering of the original geochemical data

For verification of the PCA/SVD mapping, hierarchical clustering of the original geochemical data was carried out, as well. Applying the Euclidean and Manhattan distances, the highest CC and the lowest Delta(0.5) and Delta(1.0) were obtained for the Group Average method (Table 4). The final dendrogram (Fig. 4) was constructed utilizing the Group Average method with the Euclidean distance. The clusters were denoted in consistency with the group identification in Fig. 3. As it is obvious from this figure the data structure is not so clearly organized as it was found by PCA, specially in the case of the mixed clusters I+III/1 and I+IV/1. The better clustering results of PCA are likely caused by the its noise filtration which is due to the data dimensionality reduction.

Coal type	I		II/1		II/2		III/1		III/2		IV/1		IV/2	
	x	s	x	s	x	s	x	s	x	s	x	s	x	s
W^a	1.78	0.549	4.83	1.813	9.05	1.037	10.40	2.846	19.62	3.556	10.71	3.427	17.19	1.641
A^d	7.57	4.765	11.32	3.626	4.46	0.764	21.33	15.68	9.95	9.787	28.11	11.73	14.87	4.067
V^{daf}	26.20	9.141	7.01	1.851	5.67	3.429	30.93	5.067	21.05	3.182	31.98	14.38	34.09	3.997
Q^{daf}	34.40	1.282	33.45	0.715	32.73	0.370	30.31	2.984	30.23	1.090	25.54	1.450	25.17	1.185
HA^{daf}	0.12	0.209	0.08	0.007	0.01	0.005	33.80	27.94	0.40	0.272	74.70	7.341	0.73	0.389
R_o	0.94	0.320	4.12	0.549	5.03	0.844	1.14	0.144	1.61	0.251	1.13	0.071	2.01	0.262
C^{at}	57.13	1.746	74.82	2.321	83.52	5.548	58.99	1.257	67.22	2.238	57.86	0.928	63.80	3.444
H^{at}	38.11	2.340	20.80	1.976	12.84	4.742	29.98	1.320	24.11	1.502	26.55	2.065	20.01	1.954
N^{at}	0.90	0.095	0.86	0.080	0.86	0.050	1.22	0.202	1.17	0.104	2.96	1.180	2.35	0.674
O^{at}	3.78	0.936	3.36	1.187	2.76	0.914	9.67	0.986	7.31	1.491	12.31	1.088	13.52	1.552

x-average and s-standard deviation; their units are the same as in Table 1.

Table 3 Basic statistics of the selected coal groups.

Clustering Method	Distance	Delta(0.5)	Delta(1.0)	CC
Single Linkage	Euclidean	1.0961	1.2440	0.6306
Complete Linkage	Euclidean	0.3931	0.69	0.6912
Group Average	**Euclidean**	**0.1959**	**0.2445**	**0.7682**
Simple Average	Euclidean	0.2342	0.2911	0.6916
Centroid	Euclidean	0.6431	0.8526	0.6625
Median	Euclidean	0.5672	0.5884	0.5254
Ward's Min. Variance	Euclidean	0.8983	0.9081	0.5789
Single Linkage	Manhattan	1.4639	1.7252	0.5551
Complete Linkage	Manhattan	0.4219	0.4816	0.6819
Group Average	**Manhattan**	**0.2476**	**0.2981**	**0.7109**
Simple Average	Manhattan	0.2636	0.3212	0.6760
Centroid	Manhattan	0.8672	1.0136	0.6848
Median	Manhattan	0.4905	0.5295	0.6058
Ward's Min. Variance	Manhattan	0.8745	0.8905	0.6229

Table 4 Hierarchical clustering of the standardized geochemical data.

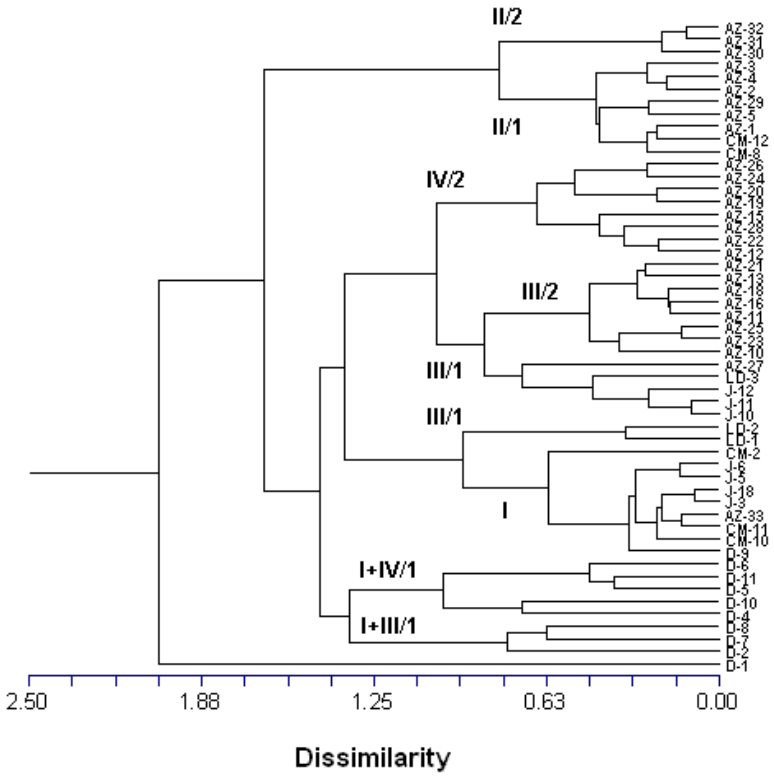

Fig. 4 Average group dendrogram of the coal samples using the Euclidean distance.

4 Conclusion

The SVD-based Principal Component Analysis of the standardized data has revealed the three principal components which explain 85.2 % of the variance. The relations among the geochemical variables were recognised by the components weights plot and also confirmed by Factor Analysis. The samples were clustered into seven groups in accordance with the intensity of thermal and oxidation changes in coal matter. The PCA clustering was compared with the Average Group dendrogram of the same data. Due to the data noise filtration properties PCA provides the better resolved cluster of the coal samples.

Acknowledgment

Author kindly thanks to Pavel Praks for the programming of SVD in MATLAB and to professor Zdeněk Klika (both from VSB-Technical University Ostrava) for providing the testing data of altered coals. This work was supported by the Ministry of Education, Youth and Sport of the Czech Republic (MSM 6198910016).

References

[1] P. Comon: "Independent Component Analysis, a new concept?", *Signal Process.*, Vol. 36, (1994), pp. 287–314.

[2] H. Attias: "Independent factor analysis", *Neural Comput.*, Vol. 11, (1998), pp. 803–851.

[3] C.M. Bishop, M. Svensén and C.K.I. Williams: GTM: "The generative topographic mapping", *Neural Comput.*, Vol. 10, (1998), pp. 215–234.

[4] P. Geladi and B.R. Kowalski: "Partial least square regression: A tutorial", *Anal. Chim. Acta*, Vol. 185, (1986), pp. 1–17.

[5] P. Geladi: „Chemometrics in spectroscopy. Part 1. Classical chemometrics", *Spectrochim. Acta Part B*, Vol. 58, (2003), pp. 767–782.

[6] E.R. Malinowski: *Factor Analysis in Chemistry*, 2nd ed., John Wiley & Sons, New York, 1991.

[7] M.E. Wall, A. Rechtsteiner and L.M.M. Rocha: "Singular value decomposition and principal component analysis", In: D.P. Berrar, W. Dubitzky and M. Granzow (Eds): *A Practical Approach to Microarray Data Analysis*, Kluwer, Norwell, MA, 2003.

[8] M.W. Berry, Z. Drmač and E.R. Jessup: "Matrices, Vector Spaces, and Information Retrieval", *Siam. Rev.*, Vol. 41, (1995), pp. 335–362.

[9] P. Praks P., J. Dvorský and V. Snášel: *Latent Semantic Indexing for Image Retrieval Systems. SIAM Conference on Applied Algebra, July 15-19, Williamsburg, 2003*, http://www.siam.org/meetings/la03/proceedings/Dvorsky.pdf

[10] Safavi and H. Abdollahi: "Thermodynamic characterization of weak association equilibria accompanied with spectral overlapping by a SVD-based chemometric method", *Talanta*, Vol. 53, (2001), pp. 1001–1007.

[11] Z. Klika and J. Kraussová: "Properties or Altered Coals Associated with Carboniferous Red Beds in the Upper Silesian Coal Basin and their Tentative Classification", *Int. J. Coal. Geology*, Vol. 22, (1993), pp. 217–235.

[12] B.K. Lavine: "Clustering and Classification of Analytical Data", In: R.A. Meyers (Ed.): *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, Chichester, 2000.

[13] P. Mather: *Computational Methods of Multivariate Analysis in Physical Geography*, John Wiley & Sons, New York, 1976.

[14] M. Kurková, Z. Klika, Ch. Kliková and J. Havel: "Humic acids from oxidized coals. I. Elemental composition, titration curves, heavy metals in HA samples, nuclear magnetic resonance spectra of HA and infrared spectroscopy", *Chemosphere*, Vol. 54, (2004), pp. 1237–1245.